Recherche de motifs spatio-temporels de cas atypiques pour le trafic routier urbain

Marc Joliveau*, Florian De Vuyst*

*Laboratoire Mathématiques Appliquées aux Systèmes, ECP
Grande Voie des Vignes
92295 Chatenay-Malabry cedex, France.
marc.joliveau@ecp.fr, florian.de-vuyst@ecp.fr

Résumé. Un large panel de domaines d'application utilise des réseaux de capteurs géoréférencés pour mesurer divers évènements. Les séries temporelles fournies par ces réseaux peuvent être utilisées dans le but de dégager des connaissances sur les relations spatio-temporelles de l'activité mesurée.
Dans cet article, nous proposons une méthode permettant d'abord de détecter des situations atypiques (au sens de l'occurrence) puis de construire des motifs spatio-temporels relatant leur propagation sur un réseau. Le cas étudié est celui du trafic routier urbain. Notre raisonnement se fonde sur l'application de la méthode Space-Time Principal Component Analysis (STPCA) et de la combinaison entre l'information mutuelle et l'algorithme Isomap.
Les résultats expérimentaux exécutés sur des données réelles de trafic routier démontrent l'efficacité de la méthode introduite à identifier la propagation de cas atypiques fournissant ainsi un outil performant de prédiction de la circulation intraday à court et moyen terme.

1 Introduction

Durant les dernières décennies, l'utilisation de réseaux de capteurs a été largement développée pour mesurer et observer l'évolution de systèmes complexes à forte dynamique. Les applications sont par exemple le trafic routier, le transport d'énergie, les processus d'entreprise et la météorologie. Dégager des liens de corrélations dans un tel réseau à travers le temps permet, par exemple, d'établir des prévisions probabilistes à court ou moyen terme. Dans ce qui suit, on suppose que les capteurs, effectuant des mesures sur le trafic routier urbain, sont fixes et géoréférencés. Un graphe de connexion logique représente les échanges ou les causalités directes possibles entre ces différents lieux géographiques. Le graphe est supposé connu.
A l'aide d'un outil d'estimation efficace, on peut prédire le comportement usuel du trafic devant chaque capteur. Cependant, lorsque la circulation est atypique, au sens de l'occurrence, la qualité des prévisions s'en retrouve considérablement affectée. Nous proposons d'identifier des motifs spatio-temporels de propagation de ces cas atypiques ayant pour objectif d'aider à prévoir les conséquences d'un évènement inhabituel sur l'intégralité du réseau.
Les motifs se réfèrent généralement à des structures répétitives sur le graphe sous-jacent dans

l'espace et le temps. Des motifs décrivant des changements dans l'espace et le temps sont qualifiés de motifs spatio-temporels (*spatiotemporal patterns*). La notion de motifs spatio-temporels apparaît dans différents secteurs scientifiques tels que les géosciences (Knopoff et al. (2001)), la météorologie (Tourre et al. (1999), Imfeld (2000)) ou l'écologie (Weigand et al. (1998)). Généralement, on utilise des techniques d'analyse et de fouilles de données spatio-temporelles discrètes pour identifier ces motifs dans de grands ensembles (Tsoukatos et Gunopulos (2001), Bittner (2001)).
Dans cet article, nous proposons dans un premier temps un outil permettant de détecter les comportements atypiques. Nous introduisons ensuite une méthode fondée sur la combinaison de l'information mutuelle (Shannon et Weaver (1963)) et de l'algorithme Isomap (Tenenbaum et Langford (2000)) calculant une première version des motifs de propagation que nous tentons d'améliorer par la suite.
Les tests expérimentaux sont effectués sur des données réelles de trafic routier intra-urbain. Ces données nous ont été fournie par l'INRETS dans le cadre du projet CADDY (`http://norma.ecp.fr/wikimas/Caddy`) de l'ACI Masse de données 2003.

2 Détection de cas atypiques

2.1 Définition d'une variable d'état de circulation continue

Dans Joliveau et De Vuyst (2007), nous avons proposé une adaptation de la méthode Space-Time Principal Component Analysis (STPCA) à un ensemble de données incomplètes calculant des estimations de séries temporelles définies pour chaque instant de mesure. L'utilisation de cette méthode sur nos données de trafic intra-urbain nous a permis de dégager un ensemble complet de données de débit et de taux d'occupation provenant d'un réseau de capteur. Le débit moyen de véhicule (nombre de véhicules/instant) correspond à la quantité de véhicules étant passés dans la zone d'activité du capteur lors du dernier intervalle de mesure. Le taux d'occupation, exprimé en pourcentage symbolise quant à lui la densité de la circulation. Plus le taux d'occupation est élevé, plus la circulation est dense. Une première difficulté est de combiner ces deux informations afin de posséder une variable ayant un sens pour l'état du trafic.
La loi fondamentale du trafic provenant de la théorie du transport indique la relation entre le flot et la densité des véhicules sur une route. A partir du diagramme représentatif de cette loi, nous proposons une nouvelle variable E continue sur $[0, 1]$ nous informant sur l'état de circulation à un capteur et un instant donné. Cette variable combine à la fois les informations de débit et de taux d'occupation et nous apporte une certaine intelligibilité sur le trafic (0 symbolisant un trafic inexistant, 1 un trafic congestionné à débit nul et 0.5 une circulation optimale avec débit maximal).
Sur la figure 1 on peut voir les représentations de séries temporelles journalières d'état de circulation obtenues pour deux capteurs choisis aléatoirement. Nous avons également illustré les séries de débit et de taux associées.

2.2 Estimation du comportement moyen et détection de situations inhabituelles

La méthode STPCA définie dans De Vuyst et Joliveau (2007) offre un moyen de résumer efficacement des séries temporelles tout en conservant la majeure partie de l'information au

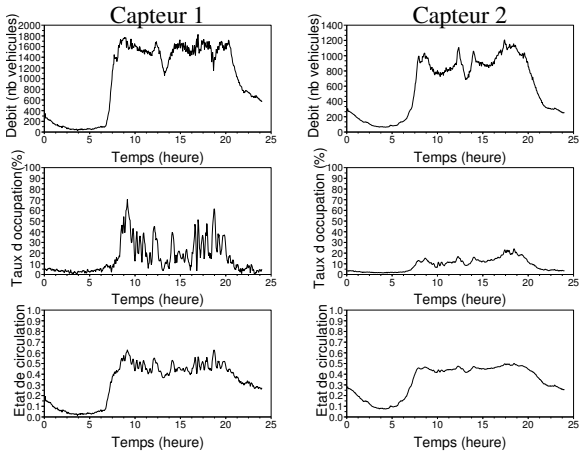

FIG. 1 – *Exemple de séries temporelles d'état de circulation pour* 2 *capteurs (ligne3). Les séries de débit (ligne 1) et de taux d'occupation (ligne 2) correspondantes sont également illustrées sur la figure.*

sens de l'énergie. L'énergie est obtenue par la trace des matrices de corrélation $\mathbf{M}$ définie par :

$$tr(\mathbf{M}) = ||\mathbf{M}||_2 = \sum_{i=1}^{N} \lambda_i(\mathbf{M})$$

où $\lambda_i(\mathbf{M})$ représente la i-ème valeur propre de $\mathbf{M}$, et N le nombre de vecteurs propres.
Le principe de la STPCA est de procéder simultanément à une analyse en composantes principales dans les deux dimensions spatiales et temporelles. Les paramètres de réduction de dimension K et L sont réglés en fonction de l'énergie capturée respectivement par les modes principaux spatiaux et temporels représentatifs de l'activité usuelle dans l'ensemble de données. Les figure 2.(a) et 2.(b) illustrent respectivement l'énergie cumulée par les premiers modes sur

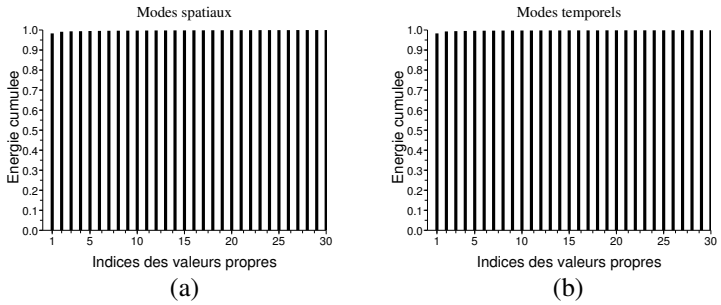

FIG. 2 – *Énergie cumulée des valeurs propres (a) spatiales et (b) temporelles.*

les dimensions spatiale et temporelle lorsqu'on applique la STPCA aux données d'état. Sur

chacune de ces deux dimensions le premier mode capture plus de 98% de l'énergie ! En outre, les $K = 4$ premiers modes spatiaux (sur 112 possibles) ainsi que les $L = 6$ premiers modes temporels (sur 480 possibles) contiennent à eux seuls plus de 99.5% de l'énergie dans les deux dimensions. Une telle proportion de l'énergie suffit à reproduire très fidèlement les comportements les plus fréquents du trafic.
Nous proposons donc d'estimer les données par STPCA en paramétrant K et L de manière à capturer la majorité de l'information au sens de l'énergie (dans nos expériences nous fixons le seuil minimal à 99.5% de l'énergie). Pour détecter un comportement atypique, il suffit de comparer la valeur réelle à son estimée par STPCA. Si l'écart entre ces deux valeurs est élevé, cela signifie que l'activité actuelle au capteur n'est pas représentative de la situation usuelle. On détecte alors un comportement atypique au sens de l'occurrence.
Notre raisonnement est illustré sur la figure 3. On remarque que, les deux courbes étant très

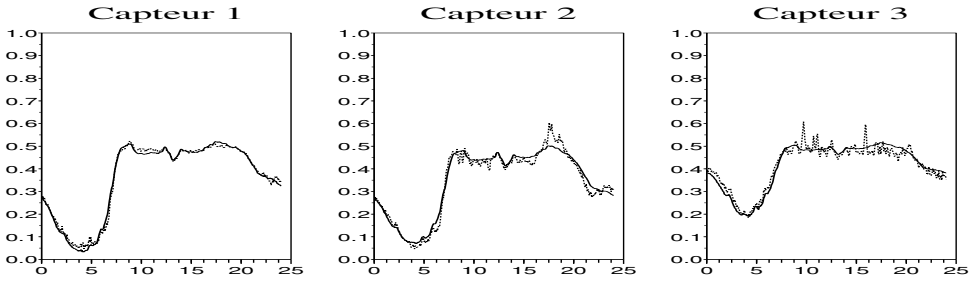

FIG. 3 – *Détection de comportement atypique en utilisant la STPCA. Les données réelles saisies par les capteurs sont représentées en pointillé, les estimées par une ligne en noir.*

proches, la série issue du capteur 1 est représentative du comportement usuel journalier à cette localisation. En ce qui concerne les deux autres capteurs, on peut identifier certaines périodes où l'activité est plus atypique. Pour le capteur 2 par exemple, on remarque que les valeurs mesurées entre $17h$ et $20h$ sont largement supérieures à leur estimation. Ces mesures relèvent un trafic surchargé, inhabituel à cette période de la journée. Il ne faut cependant pas confondre évènement atypique et congestion. Par exemple, bien que le capteur 3 illustre certaines courtes périodes de congestion qui ne sont pas habituelles (vers $9h$ et $16h$), on peut également observer un écart entre les deux courbes en fin d'après-midi. L'estimation nous indique que le trafic est habituellement chargé sur cette période (l'état de circulation est supérieur à 0.5). Or, les mesures réelles relèvent une activité plus fluide que d'habitude. Dans ce cas, le comportement atypique détecté correspond à une fluidité inhabituelle du trafic.
Dans le cadre où on cherche à établir des prévisions sur le trafic à court ou moyen terme, l'utilisation de la STPCA sur la variable E est un outil très pratique. Si l'écart entre la valeur mesurée par un capteur et son approximation est faible, on se réfère à la série estimée par l'algorithme STPCA pour établir notre prévision. Dans le cas contraire, on détecte un comportement atypique. On s'appuiera alors sur des motifs spatio-temporels de propagation pour prédire la circulation.

3 Recherche de motifs spatio-temporels

Les motifs intraday que nous cherchons à identifier ont pour but de nous aider à faire des prévisions sur le trafic. Ceux-ci se focalisent plus particulièrement sur l'anticipation de la propagation d'une situation atypique à travers le réseau. La propagation d'une situation atypique sur un réseau est une notion de voisinage locale en espace et en temps indiquant sur quels points du réseau on observe un comportement atypique suite à la réalisation d'un évènement inhabituel à l'instant précédent.

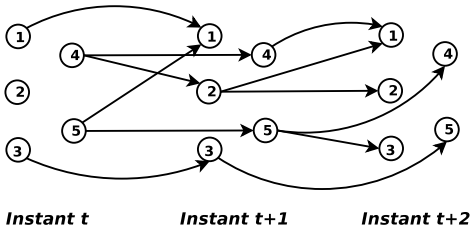

FIG. 4 – *Exemple d'un motif spatio-temporel se propageant sur* 3 *instants successifs.*

La figure 4 représente un exemple de motifs spatio-temporels de propagation sur un petit réseau de 5 capteurs au cours de 3 instants successifs. Les motifs peuvent être représentés par une chaîne ou un graphe décomposable en niveau, dans lequel chaque niveau correspond à un instant de mesure, chaque sommet symbolise un capteur à un instant donné, et un lien entre deux sommets représente la propension de propagation d'une situation atypique d'une période à la suivante. Les arcs peuvent être pondérés par la tendance de propagation. Sur cet exemple de motif, la détection d'une circulation inhabituelle au capteur 4 en t nous incite à anticiper une situation atypique aux capteurs 2 et 4 à la période suivante.
Deux outils principaux sont utilisés pour identifier les motifs spatio-temporels de propagation : l'information mutuelle (Shannon et Weaver (1963)) et l'algorithme Isomap (Tenenbaum et Langford (2000)).

3.1 L'information mutuelle

L'information mutuelle est tirée de la théorie des probabilités et de la théorie de l'information. Cette quantité mesure la dépendance statistique entre deux variables. L'information mutuelle mesurant la quantité d'information apportée en moyenne par une réalisation d'un évènement X sur les probabilités de réalisation d'un évènement Y, et, en considérant qu'une distribution de probabilité représente notre connaissance d'un phénomène aléatoire, on peut mesurer l'absence d'information en utilisant l'entropie de Shannon (Shannon et Weaver (1963)) de cette distribution. Ainsi, l'information mutuelle est donnée par :

$$I(X,Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$$

où $H(X)$ et $H(Y)$ mesurent l'entropie respective des évènements X et Y, et $H(X,Y)$ l'entropie croisée entre ces évènements.
L'information mutuelle est comprise entre 0 et 1. Plus sa valeur est élevée, plus les évènements sont liés, une information mutuelle de 0 étant synonyme d'évènements indépendants.
L'information mutuelle est utilisée dans une grande variété de domaines comme la fouille de données, l'imagerie ou la biologie moléculaire pour déterminer des relations dans le comportement de variables. Dans Liang et al. (1998) par exemple, les auteurs utilisent l'information mutuelle afin de détecter des liens d'inférence permettant de dégager des connaissances sur la régulation des gènes.

3.2 Isomap

Isomap est une méthode introduite par Tenenbaum et Langford (2000) dont l'objectif est d'identifier la structure cachée dans des observations multivariées de grandes dimensions. Le principe de cette méthode est de projeter les points d'un ensemble de données dans un espace de plus faible dimension. L'avantage d'Isomap est que contrairement aux méthodes classiques telles que l'analyse en composantes principales (ACP) ou l'échelonnement multidimensionnel (*multidimensional scaling* - MDS), l'algorithme est capable de découvrir des relations non linéraires régissant certaines observations complexes. Une fois le plongement (ou *embedding*) calculé, la distance entre deux points sur celui-ci est représentative de la similitude globale entre les évènements mesurés par ces points.
Isomap se décompose en trois étapes définies de manière détaillée dans le tableau 1. Lors de la première étape, on cherche à déterminer pour chaque élément les points qui lui sont le plus proche en fonction d'une distance $d_x(i,j)$ définie entre une paire de points i et j dans l'espace d'origine X. On peut, au choix, utiliser un rayon de distance préfixé ϵ ou calculer les K plus proches voisins de chaque élément. Cette relation de voisinage permet de construire un graphe G où un arc, pondéré par la distance $d_x(i,j)$, relie chaque point i à l'ensemble de ses voisins j.
Une fois le graphe G déterminé, les distances géodésiques entre chaque paire de point de l'ensemble sont calculées en déterminant les valeurs $d_G(i,j)$ des plus courts chemins entre chaque paire de sommet de G. Dans le résumé de la méthode, nous proposons d'utiliser l'algorithme de Floyd de complexité polynomiale ($O(n^3)$, n étant le nombre de sommets).
La dernière étape d'Isomap consiste à appliquer un MDS classique à la matrice des distances du graphe $\mathbf{D}_G = \{d_G(i,j)\}_{ij}$ afin de construire un plongement dans un espace $\mathbf{Y}$ de dimension d qui préserve au mieux l'estimation de la géométrie intrinsèque de l'espace. On, introduit l'opérateur τ convertissant les distances en produits internes qui caractérisent uniquement la géométrie des données dans une forme compatible à une optimisation efficace.

$$\tau(\mathbf{D}) = -1/2 \ \mathbf{H}\, \mathbf{S_D}\, \mathbf{H}$$

où $\mathbf{S_D}$ est la matrice des distances au carré $\{(\mathbf{S_D})_{ij} = \mathbf{D}_{ij}^2\}$, et $\mathbf{H}$ représente la "matrice de centrage" $\{\mathbf{H}_{ij} = \delta_{ij} - 1/N\}$, N étant le nombre d'observations.
Les coordonnées des vecteurs $\mathbf{y}_i$ des points de $\mathbf{Y}$ sont choisies en minimisant une fonction de coût :

$$\min \|\tau(\mathbf{D}_G) - \tau(\mathbf{D}_Y)\|_F \tag{1}$$

TAB. 1 – *Algorithme de la méthode Isomap.*

Étape	
1 Construction du graphe de voisinage	*Définir le graphe G en connectant les points i et j si $d_x(i,j)$ est plus petit que ϵ ou si j est un des K plus proches voisins de i relativement à d_x. Pondérer les arcs par la valeur $d_x(i,j)$.*
2 Calcul des plus courts chemins	*Initialiser $d_G(i,j) = d_x(i,j)$ si i et j sont reliés par un arc, $d_G(i,j) = \infty$ sinon. Ensuite pour $k = 1 \ldots N$ remplacer $d_G(i,j)$ par $\min\{d_G(i,j), d_G(i,k) + d_G(k,j)\}$. La matrice des valeurs finales $\mathbf{D}_G = \{d_G(i,j)\}$ contient les distances des plus courts chemins entre chaque paire de points de G.*
3 Construction du plongement	*Soit λ_p la p-ième valeur propre (triées en ordre décroissant) de la matrice $\tau(\mathbf{D}_G)$, et $\mathbf{v}_p^i$ la i-ème composante du p-ième vecteur propre. On affecte à la p-ième composante des coordonnées du vecteur $\mathbf{y}_i$ de dimension d la valeur $\sqrt{\lambda_p}\mathbf{v}_p^i$.*

où $\mathbf{D}_Y$ représente la matrice des distances euclidiennes $\{d_y(i,j) = \|\mathbf{y}_i - \mathbf{y}_j\|\}$ et $\|\mathbf{A}\|_F$ la norme de Frobénius de la matrice $\mathbf{A}$ donnée par : $\sqrt{\sum_{i,j} \mathbf{A}_{ij}^2}$.
Le minimum global du problème de minimisation (1) est obtenu en affectant les d premiers vecteurs propres (triés dans l'ordre décroissant) de la matrice $\tau(\mathbf{D})$ aux coordonnées de $\mathbf{y}_i$.

3.3 Construction des motifs spatio-temporels

3.3.1 Calcul de motifs d'origine

Afin de construire une première version de motifs spatio-temporels de propagation de comportement atypique dans un réseau, nous appliquons l'algorithme Isomap en utilisant l'information mutuelle comme mesure de distance $d_x(i,j)$ sur l'ensemble des données d'origine.
L'information mutuelle ne se calculant que dans un cadre discret et la variable étudiée étant continue sur $[0,1]$, nous décidons de la symboliser pour la discrétiser. Pour cela, nous proposons sept états quantifiés de circulation :

- "c" : calme ; peu de véhicules circulant, circulation nocturne typique ;
- "TGC" : Tendance à la grande circulation ; état intermédiaire entre c et GC avec une tendance à l'augmentation du trafic ;
- "rc" : Retour au calme ; état intermédiaire entre c et GC avec une tendance à la diminution du trafic ;
- "GC" : Grande circulation ; état correspondant à une circulation quasi-optimale, un débit très élevé mais pas d'engorgement ;
- "$S1$" : Saturation de niveau 1 ; correspond à une petite congestion, le débit de véhicule diminue alors que le taux d'occupation augmente ;
- "$S2$" : Saturation de niveau 2 ; correspond à une congestion assez grosse, les véhicules circulent lentement et le trafic est très chargé ;
- "$S3$" : Saturation de niveau 3 ; le trafic est quasiment ou totalement saturé et les véhicules à l'arrêt ;

La transformation de la variable E à sa version symbolique est appliquée à l'aide d'intervalles ainsi que, pour certains états, par le signe de la dérivée de la série temporelle à l'instant mesuré. Le tableau 2 renvoie les seuils utilisés pour symboliser les séries. Cette discrétisation a

TAB. 2 – *Calcul des symboles à partir des données E.*

Symbole	Valeur de E en t	Signe de la dérivé en t
c	$E < 0.2$	/
rc	$0.2 \leq E < 0.45$	négatif
TGC	$0.2 \leq E < 0.45$	positif
GC	$0.45 \leq E < 0.52$	/
S1	$0.52 \leq E < 0.6$	/
S2	$0.6 \leq E < 0.7$	/
S3	$E \geq 0.7$	/

été choisie dans le but de calculer les dépendances entre les capteurs en fonction d'informations intelligibles et utiles pour un utilisateur.
L'application d'Isomap avec l'information mutuelle sur les données journalières en différenciant chaque période nous fournit autant de plongements que d'instants de mesure. Sur chaque plongement, les distances entre les points (chaque point symbolisant un capteur) sont représentatives des similitudes de comportement entre les capteurs à une période donnée.
On norme les distances sur chaque plongement en fonction de la plus grande distance sur le plongement. Ainsi, pour chaque instant on possède une distance de similitude comprise entre 0 et 1. Nous construisons une première version de motifs spatio-temporels à partir des plongements fournis par Isomap. Pour chaque période t, pour chaque capteur i, on identifie les capteurs les plus proches de i sur le plongement correspondant à la période $t+1$ et on relie ces deux couples capteur-instant par un arc. De cette manière, on obtient un graphe du même type que sur la figure 4 représentant les premiers motifs de propagation de cas atypiques.
La notion de capteur le plus proche peut être définie par un rayon de distance minimale ϵ ou par l'algorithme des K plus proches voisins. Dans nos expériences, nous avons combiné les deux approches en choisissant les K plus proches voisins dans un rayon de ϵ.

3.3.2 Amélioration des motifs

Suite à la combinaison d'Isomap et de l'information mutuelle, on dispose de premiers motifs spatio-temporels de propagation pouvant être représentés par une chaîne reliant les capteurs d'un réseau entre les périodes de mesure. Nous proposons de pondérer les arcs de cette chaîne par la tendance de propagation du caractère atypique.
Pour cela, on simule le déroulement du temps en détectant les cas atypiques par le procédé expliqué dans la section 2 et on calcule les probabilités $P(A_{j,t+1}|A_{i,t})$ qu'il y ait un évènement atypique en j à l'instant $t+1$ sachant qu'il y a un évènement atypique en i à l'instant t. Cette valeur considérée comme la tendance de propagation sert à pondérer les arcs (i,j) de la chaîne des motifs.
A partir de ces tendances de propagation deux types d'amélioration peuvent être appliquées aux motifs. D'un coté, on peut retirer les capteurs dont la tendance de propagation est trop faible, c'est-à-dire inférieure au seuil de probabilité final σ_{pf}.

D'un autre coté, on cherche également à agréger aux motifs certains capteurs qui seraient susceptibles d'augmenter la qualité des prédictions. Dans cet objectif, nous définissons les notions de voisins directs et indirects :

Les *voisins directs* d'un couple capteur-temps (i, t) sont les couples capteurs-temps $(j, t+1)$ liés au premier couple dans la chaîne des motifs ;

Les *voisins indirects* de (i, t) sont les voisins directs issus des voisins directs de (i, t) dont la tendance de propagation est élevée, c'est-à-dire supérieure au seuil de voisinage σ_v.

Le calcul des voisins indirects est itératif. Si un voisin indirect renvoie une tendance de propagation suffisamment forte, on l'ajoute aux motifs puis on le considère comme un voisin direct lors de l'itération suivante.

4 Expériences numériques

Les expériences ont été réalisées à partir du jeu de données réelles de trafic routier urbain fourni par l'INRETS dans le cadre du projet CADDY. Le jeu de données utilisé considère les valeurs de débit et de taux d'occupation issues de 112 capteurs sur 100 jours de semaine. Les mesures ont lieu toutes les 3 minutes. A partir de ces données, on déduit les séries temporelles d'état de circulation pour la même quantité de jour et de capteur à une échelle temporelle identique.

4.1 Expériences sur l'ensemble des données

Un taux d'échantillonage des séries temporelles fixé à 3 minutes nous permet de tester notre méthode d'identification de motifs à court terme. Dans nos expériences, différents éléments de mesure vont être calculés :

- La moyenne de faux positifs (FP) par alarme. Cette valeur est calculée par le ratio entre le nombre total de faux positifs (anticipation erroné d'une situation atypique) et le nombre total de situations atypiques.
- La moyenne de faux négatifs (FN) par alarme. Cette valeur est calculée par le ratio entre le nombre total de faux négatifs (évènement atypique non anticipé) et le nombre total de situations atypiques.
- L'efficacité moyenne. Cette variable calculée par le ratio entre le nombre de situations atypiques prévues et le nombre total de situations atypiques réelles renvoie le pourcentage de cas atypiques anticipés à bon escient.

Ces trois mesures sont calculées en simulant des prévisions en temps réel sur l'intégralité des données à partir des motifs identifiés par la méthode. Le tableau 3 indique les résultats obtenus par la prévision de situations atypiques à l'aide des motifs de propagation calculés selon différentes valeurs des paramètres σ_{pf} et σ_v. La colonne "voisins directs" correspond aux résultats obtenus à l'aide de la première version des motifs tandis que la colonne "voisins indirect" se réfère à la méthode avec les améliorations. On remarque que la solution basée sur l'utilisation exclusive des voisins directs, déjà de bonne qualité, permet de prévoir jusqu'à $80,5\%$ des situations atypiques avec un taux de faux positifs acceptable. Bien que notre but soit de prévoir la propagation de cas atypiques au sens de l'occurrence, il nous faut distinguer les éléments atypiques reproductibles des éléments atypiques trop ponctuels. Le réglage du seuil σ_{pf} permet de gérer ce compromis. De son coté, l'ajout des voisins indirects augmente la précision

Avec $\sigma_{pf} = 0$

	Voisins directs	**Voisins Indirects**		
		$\sigma_v = 0$	$\sigma_v = 0.7$	$\sigma_v = 0.9$
moyenne de FP par alarme	1.28	2.07	1.87	1.49
moyenne de FN par alarme	0.2	0.17	0.18	0.19
Efficacité moyenne (en %)	80.5	82.7	82.2	81.0

Avec $\sigma_{pf} = 0.25$

	Voisins directs	**Voisins Indirects**		
		$\sigma_v = 0$	$\sigma_v = 0.7$	$\sigma_v = 0.9$
moyenne de FP par alarme	0.32	0.41	0.39	0.31
moyenne de FN par alarme	0.21	0.19	0.19	0.20
Efficacité moyenne (en %)	79.5	81.0	80.6	80.0

Avec $\sigma_{pf} = 0.75$

	Voisins directs	**Voisins Indirects**		
		$\sigma_v = 0$	$\sigma_v = 0.7$	$\sigma_v = 0.9$
moyenne de FP par alarme	0.08	0.08	0.08	0.08
moyenne de FN par alarme	0.43	0.43	0.43	0.43
Efficacité moyenne (en %)	57.0	57.1	57.1	57.0

TAB. 3 – *Résultats obtenus sur l'ensemble des données (*100 *jours) avec un échantillonnage de* 3 *minutes pour différentes valeurs du seuil de probabilité final (σ_{pf}) et du seuil de voisinage (σ_v).*

au niveau de l'efficacité de prédiction (permettant un gain de 2.2% avec $\sigma_{pf} = \sigma_v = 0$). Il faut cependant le paramétrer de manière à ne pas trop augmenter le nombre de faux positifs engendrés.

4.2 Apprentissage sur un échantillon

	Entraînement sur 50 **jours**	**Test sur** 100 **jours**
moyenne de FP par alarme	0.38	0.40
moyenne de FN par alarme	0.19	0.22
Efficacité moyenne (en %)	81.1	78.3

TAB. 4 – *Résultats obtenus à court terme (*3 *minutes) en entraînant l'algorithme sur* 50 *jours puis en le testant sur l'ensemble des dates (avec $\sigma_v = 0.7$ et $\sigma_{pf} = 0.25$).*

Le but de notre travail étant de déterminer des motifs spatio-temporels afin de réaliser des prévisions en temps réel, nous aimerions estimer les facultés d'apprentissage de notre méthode. Pour cela, nous sélectionnons aléatoirement un échantillon de nos données et nous procédons aux différentes étapes de notre méthode à partir de cet échantillon (détermination des modes principaux, calcul de l'information mutuelle, détermination des premiers motifs avec Isomap, amélioration des motifs). Finalement, on teste les capacités de prédiction des motifs extraits de l'échantillon sur l'intégralité des données.
Le tableau 4 illustre les résultats obtenus en utilisant un échantillon de 50 jours et en fixant les valeurs des paramètres à $\sigma_v = 0.7$ et $\sigma_{pf} = 0.25$. L'efficacité de prédiction reste élevée (78.3%

par rapport à 80.6% lorsque les motifs sont calculés sur l'intégralité des données) alors que la quantité de faux positifs demeure acceptable.
La figure 5 illustre un extrait des motifs identifiés dans ce cadre. La partie gauche de la figure

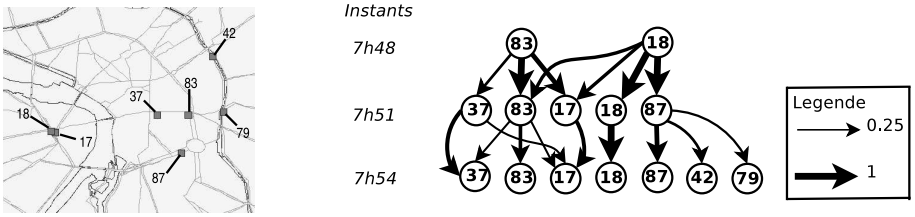

FIG. 5 – *Illustration d'un extrait des motifs spatio-temporels de propagation.*

représente la topologie du réseau, les capteurs étant représentés par des carrés, les routes par des traits gris clairs et les cours d'eau par des traits fins noirs. La partie droite illustre un morceau de la chaîne de motifs correspondant aux instants entre $7h48$ et $7h54$. L'importance de la tendance de propagation est symbolisée par l'épaisseur des flèches.

4.3 Prévisions à moyen terme

En modifiant l'échantillonnage temporel, on peut appliquer notre méthode sur des épisodes plus longs. Le tableau 5 illustre les résultats de la méthode à moyen terme (sur des épisodes de l'ordre du quart d'heure) obtenus à partir d'un échantillon de 50 jours choisis aléatoirement avec $\sigma_v = 0.7$ et $\sigma_{pf} = 0.25$.
L'efficacité est légèrement altérée mais reste tout à fait acceptable (65.6%) avec une proportion

	Entraînement sur 50 jours	**Test sur 100 jours**
moyenne de FP par alarme	0.39	0.41
moyenne de FN par alarme	0.32	0.34
Efficacité moyenne (en %)	68.2	65.6

TAB. 5 – *Résultats obtenus à moyen terme (15 minutes) en entraînant l'algorithme sur 50 jours puis en le testant sur l'ensemble des dates (avec $\sigma_p = 0.7$ et $\sigma_{pf} = 0.25$).*

de faux positifs similaire aux prévisions à court terme.

5 Conclusion

Dans ce papier, nous avons proposé une méthode de génération de motifs spatio-temporels de situations atypiques. Cette méthode s'applique principalement à des données issues d'un réseau de capteurs fixes géoréférencés. Nous utilisons l'algorithme STPCA comme outil de prévision du comportement usuel de séries temporelles issues d'un réseau de capteurs, nous permettant également de détecter les situations atypiques. Nous avons également introduit une méthode d'identification de motifs spatio-temporels de propagation de situations atypiques fondée sur la combinaison d'Isomap et de l'information mutuelle. Ces motifs ont pour fonction

d'aider à réaliser des prévisions à court et moyen terme sur le réseau.
Les expériences numériques réalisées sur un ensemble de données réelles de trafic routier ont su démontrer la qualité des motifs en terme de prédiction et les facultés d'apprentissage de la méthode à court et moyen terme.

Références

Bittner, T. (2001). Rough sets in spatio-temporal data mining. In *Workshop on Temporal, Spatial and Spatio-Temporal Data Mining*, pp. 89–104.

De Vuyst, F. et M. Joliveau (2007). Space-time principal component analysis for multivariate time series. *MAS Lab. Research Rapport 0604*, `http://www.mas.ecp.fr/download_publication.php?type=techreport&referenc%e=0604&file=url`.

Imfeld, S. (2000). Time, points and space - towards a better analysis of wildlife data in gis. In *Dissertation*. University of Zürich.

Joliveau, M. et F. De Vuyst (2007). Space-time summarization of multisensor time seires. case of missing data. In *Int. Workshop on Spatial and Spatio-temporal data mining, IEEE ICDM*. To appear.

Knopoff, L., A. Gabrielov, et M. Ghils (2001). *IMA Workshops on Spatio-Temporal Patterns in the Geosciences*. University of Minnesota.

Liang, S., S. Fuhrman, et R. Somogyi (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing*, Volume 3, pp. 18–29.

Shannon, C. et W. Weaver (1963). *The Mathematical Theory of Communication*. University of Illinois Press.

Tenenbaum, J. et J. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 2319–2323.

Tourre, Y., B. Rajagopalan, et Y. Kushnir (1999). Dominant patterns of climate variablility in the atlantic ocean during the last 136 years. *Journal of Climate 12*, 2285–2299.

Tsoukatos, I. et D. Gunopulos (2001). Efficient mining of spatio-temporal patterns. In *Proc. of* 7^{th} *Int. Symp. on Spatial and Temporal Databases*, pp. 425–442.

Weigand, T., K. Moloney, et S. Milton (1998). Population dynamics, disturbance, and pattern evolution : Identifying the fundamental scales of organization in a model ecosystem. *The American Naturalist 152*, 321–337.

Summary

A wide variety of application domains uses sensors networks to measure differents events. Times series returned by such networks can be then used to extract knowledge on activity spatio-temporal relationships.
In this paper, we introduce a method able to detect non-typical situations and then to build spatio-temporal patterns given information about their propagation in the network. Our idea is based on the application of the so-called Space-Time Principal Component Analysis (STPCA) method and the combination between mutual information and Isomap algorithm.
Experimental tests applied on real life road trafic data demonstrate the ability of the proposed method to identify spread of non-typical cases giving an accurate tool to make prediction of the circulation at short-term and mid-term.